Jurnal Sains Teknologi dan Sistem Informasi
journal.yp3a.org/index.php/satesi
Vol. 5 No. 1 (2025) 51 – 60 ISSN Media Elektronik: 2807–8152

# Penerapan Hadoop untuk Analisis Sentimen Berbasis Big Data pada Ulasan Aplikasi Transportasi *Online*

Putri Angraini Aziz[1*], Syaban Barokah Nur Ilahi[2], Sumiarni Moka[3], Adha Mashur Sajiah[4]
[1,2,3,4]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Halu Oleo, Kendari, Indonesia
Email: [1*]putriangrainiaziz27@gmail.com, [2]syabanbarokahnurilahi@gmail.com , [3]sumiarni.mk@gmail.com, [4]adha.m.sajiah@uho.ac.id

***ABSTRACT***
*The rapid growth of application-based transportation services in Indonesia has generated a large volume of user reviews that contain essential information for service development. However, significant challenges arise in processing and analyzing data on a large scale. This study utilizes Hadoop and Apache Spark technology to conduct sentiment analysis on online transportation application reviews, focusing on Gojek user reviews. The dataset comprises 1.880.112 reviews obtained from Kaggle and Google Play Store. The research method includes data preprocessing using distributed computing with Hadoop and Spark, followed by sentiment labeling based on user ratings. The sentiment analysis model used is Logistic Regression, with hyperparameter tuning through Cross Validation. The evaluation results show a model accuracy of 80%, demonstrating the model's capability in effectively classifying sentiments, supported by PySpark implementation which enables efficient training and evaluation processes despite working with large-scale datasets. Text visualization in the form of a word cloud reveals that negative sentiment is often associated with app performance and digital wallet issues, while neutral sentiment focuses more on driver services. On the other hand, positive sentiment highlights user satisfaction with the overall service. The findings of this study demonstrate the effectiveness of Hadoop in large-scale sentiment analysis processing and provide valuable insights for improving online transportation services.*

*Keywords: Hadoop, Apache Spark, Sentiment Analysis, Logistic Regression.*

**ABSTRAK**
Pertumbuhan pesat layanan transportasi berbasis aplikasi di Indonesia telah menciptakan sejumlah besar ulasan pengguna yang menyimpan informasi penting untuk pengembangan layanan. Namun, tantangan signifikan muncul dalam hal pemrosesan dan analisis data dalam skala besar. Penelitian ini memanfaatkan teknologi Hadoop dan Apache Spark untuk menganalisis sentimen ulasan transportasi online, dengan fokus pada ulasan pengguna Gojek. Dataset yang digunakan mencakup 1.880.112 ulasan yang diambil dari Kaggle dan Google Play Store. Metode penelitian meliputi *preprocessing* data menggunakan komputasi terdistribusi dengan Hadoop dan Spark, diikuti oleh pelabelan sentimen berdasarkan penilaian pengguna. Model analisis sentimen yang digunakan adalah *Logistic Regression*, dengan penyesuaian *hyperparameter* melalui *Cross Validation*. Hasil evaluasi menunjukkan akurasi model sebesar 80% yang menunjukkan kemampuan model dalam mengklasifikasikan sentimen dengan baik, didukung oleh implementasi PySpark sehingga memungkinkan proses *training* dan evaluasi dapat dijalankan secara efisien walaupun dengan dataset berukuran besar. Visualisasi teks dalam bentuk word cloud menunjukkan bahwa sentimen negatif sering kali terkait dengan kinerja aplikasi dan masalah dompet digital, sedangkan sentimen netral lebih berfokus pada layanan pengemudi. Di sisi lain, sentimen positif menekankan kepuasan pengguna terhadap layanan secara keseluruhan. Temuan dari penelitian ini menunjukkan efektivitas penggunaan Hadoop dalam pemrosesan analisis sentimen berskala besar, serta memberikan wawasan penting untuk perbaikan layanan transportasi online.

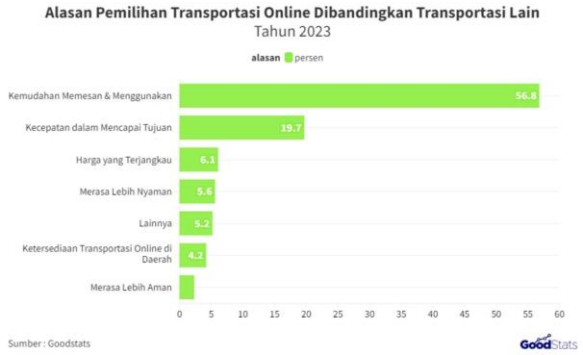
Kata Kunci: Hadoop, Apache Spark, Analisis Sentimen, *Logistic Regression.*

## 1. Pendahuluan

Transportasi *online* seperti maxim, gojek, dan grab merupakan aplikasi transportasi yang sedang berkembang pesat di Indonesia [1]. Berdasarkan survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), layanan transportasi *online* berada di peringkat ke-16 dari 22 alasan utama masyarakat menggunakan internet di Indonesia[2]. Sementara itu, menurut survey yang dilakukan

Diterima Redaksi: 19-02-2025 | Selesai Revisi: 15-03-2025 | Diterbitkan Online: 20-04-2025

Goodstats pada tanggal 8-15 Juni 2023 dengan judul "Pola Perilaku Masyarakat Indonesia Saat Menggunakan Transportasi *Online*", menunjukkan bahwa 56.8% masyarakat memilih transportasi *online* karena kemudahan pemesanan penggunaan. Selain itu, kecepatan tujuan (19.7%) dan harga terjangkau (6.1%) juga menjadi faktor penting dalam preferensi masyarakat terhadap transportasi *online* dibandingkan transportasi konvensional [3], hal ini seperti ditunjukkan pada Gambar 1.




Gambar 1. Alasan pemilihan moda transportasi online | Goodstats

Perkembangan tersebut terus berlanjut dan meningkat pada berbagai aspek transportasi, di mana transportasi yang baik dapat membantu meningkatkan perekonomian suatu daerah.

Beberapa alasan utama penggunaan transportasi daring adalah karena kemudahan akses melalui aplikasi *smartphone* yang ada, murah, dan aman [4]. Selain itu, penggunaan jasa transportasi daring lebih cenderung menggunakan konsep *on-demand business*, yaitu layanan yang didasarkan pada adanya permintaan konsumen. Dalam model bisnis ini, kepuasan dan loyalitas pelanggan menjadi kunci utama keberlanjutan layanan [5]. Karena itulah, penyedia jasa transportasi *online* perlu melakukan pemantauan dan evaluasi tingkat kepuasan pengguna terhadap kinerja sistem maupun layanan mereka. Hasil pengukuran sistem yang digunakan berupa rekomendasi bagi perusahaan agar ke depannya sistem yang digunakan dapat diimplementasikan dengan lebih baik.

Saat ini, jumlah pengguna layanan transportasi *online* semakin meningkat, yang berdampak pada melonjaknya volume ulasan pengguna. Analisis sentimen berbasis *big data* menjadi penting karena metode konvensional tidak mampu mengolah data dalam skala besar secara efisien. Dengan pemrosesan berbasis Hadoop dan Spark, analisis sentimen dapat dilakukan lebih cepat, memberikan wawasan yang dapat dimanfaatkan secara *real-time* oleh penyedia layanan transportasi online. Selain itu, hasil analisis sentimen dapat digunakan untuk mengidentifikasi masalah layanan dan meningkatkan kepuasan pelanggan[6].

Di antara berbagai platform transportasi *online* yang ada, Gojek menjadi salah satu yang paling berpengaruh dengan jumlah pengguna yang cukup besar di Indonesia [7]. Hal ini menjadikan Gojek sebagai objek studi yang ideal untuk memahami sentimen pengguna terhadap layanan transportasi *online*. Pengguna yang secara aktif memberikan ulasan dan penilaian terhadap layanan Gojek, menjadikan data yang dapat dianalisis untuk mengukur kepuasan pengguna menjadi melimpah. Untuk menganalisis data dengan volume yang cukup besar, diperlukan beberapa metode yang dapat melakukan analisis sentimen secara efektif dan efisien.

Beberapa hasil penelitian sebelumnya menunjukkan bahwa terdapat berbagai pendekatan yang dapat diterapkan dalam analisis sentimen, di antaranya *Logistic Regression, Naïve Bayes, dan Support Vector Machine (SVM)*. *Logistic Regression* merupakan algoritma yang sederhana namun efektif, dengan performa yang kompetitif pada dataset dengan distribusi kelas yang seimbang serta kemampuan menghasilkan probabilitas output yang terkalibrasi dengan baik [8][9]. Di sisi lain, *Naïve Bayes* unggul dalam hal kecepatan, tetapi serung kali kurang akurat ketika berhadapan dengan data teks yang rumit [10]. Sementara itu, SVM dikenal memiliki keunggulan dalam menangani data yang kompleks dan mampu menghasilkan klasifikasi yang akurat [11], tetapi memerlukan waktu komputasi yang tinggi ketika dataset yang dianalisis sangat besar [12]. Namun, ketika berhadapan dengan data yang sangat besar seperti ulasan aplikasi transportasi *online*, metode konvensional menghadapi masalah dalam hal efisiensi dan skalabilitas.

Pertumbuhan data yang terus berlipat ganda dari waktu ke waktu melampaui kapasitas media penyimpanan maupun sistem database tradisional, sehingga muncul istilah *Big Data*. *Big Data* merujuk pada kumpulan data yang berukuran besar, berkecepatan tinggi, kompleks, dan bervariasi, yang membutuhkan teknik serta teknologi tingkat lanjut untuk pengumpulan, penyimpanan, distribusi, manajemen, dan analisis[13]. Untuk mengatasi tantangan ini, Hadoop diperkenalkan sebagai solusi untuk menangani pemrosesan data dalam jumlah besar secara terdistribusi. Hadoop memungkinkan pembagian tugas ke beberapa node, yang dapat mengurangi waktu pemrosesan dan meningkatkan[14]. Namun, Hadoop sendiri tidak menyediakan algoritma *machine learning* seperti *logistic regression* secara langsung, sehingga diperlukan *framework* tambahan untuk menjalankan analisis sentimen.

Salah satu *framework* yang dapat bekerja di atas Hadoop adalah Apache Spark, yang menawarkan dukungan untuk *machine learning* dalam skala besar. Spark menyediakan berbagai *tools* untuk *machine learning* dan analisis sentimen, serta dapat berjalan secara terdistribusi di atas Hadoop[6]. Dengan kombinasi antara Spark dan Hadoop, pemrosesan data skala besar untuk analisis sentimen dapat dilakukan lebih cepat dan juga efisien.

DOI: 10.54259/satesi.v5i1.4051

Pada penelitian sebelumnya, Saripah Aini Pohan membahas analisis sentimen pada aplikasi Maxim. Hasil penelitian ini menunjukkan efektivitas metode *bagging* dalam meningkatkan akurasi klasifikasi ulasan pengguna terhadap aplikasi Maxim yang mana akurasinya mencapai 64% (Aini Pohan & Hasyifah Sibarani,2024)[15]. Sementara itu, penelitian Melia mengkaji analisis sentimen terhadap pengguna Gojek dan Grab di Twitter. Studi ini menunjukkan bahwa algoritma *Random Forest* mampu mencapai akurasi sekitar 76,25% dalam mengklasifikasikan sentimen pengguna Grab dan Gojek (Irawan & Nurdiawan, 2023)[16].

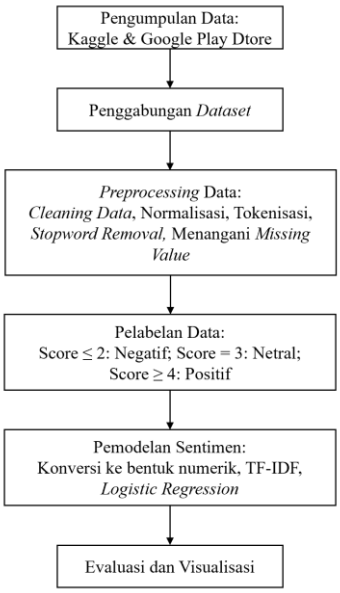
Berdasarkan latar belakang di atas, penelitian ini bertujuan untuk menjawab beberapa pertanyaan penting mengenai analisis sentimen pengguna Gojek. Pertama, bagaimana distribusi sentimen positif, negatif, dan netral pada ulasan aplikasi Gojek dapat memberikan gambaran umum terkait tingkat kepuasan pengguna. Kedua, faktor-faktor apa saja yang paling sering muncul dalam sentimen negatif, yang dapat menjadi fokus perbaikan layanan. Ketiga, sejauh mana model *Logistic Regression* dapat mengklasifikasikan sentimen dengan akurasi yang baik.

Sejalan dengan rumusan masalah tersebut, penelitian ini memiliki beberapa tujuan utama. Pertama, menganalisis sentimen pengguna aplikasi Gojek dengan menggunakan metode *Logistic Regression* berbasis Hadoop dan Spark untuk mendapatkan wawasan mengenai persepsi pengguna terhadap layanan. Kedua, mengidentifikasi faktor-faktor utama yang berkontribusi terhadap sentimen negatif dalam ulasan pengguna, sehingga dapat menjadi dasar rekomendasi untuk perbaikan layanan. Ketiga, mengevaluasi performa model *Logistic Regression* dalam klasifikasi sentimen ulasan pengguna dengan menggunakan berbagai metrik evaluasi seperti akurasi, *precision, recall,* dan *F1-score* untuk memastikan keandalan hasil analisis.

Untuk mencapai tujuan tersebut, penelitian ini menggunakan pendekatan berbasis Big Data dengan Hadoop sebagai sistem penyimpanan dan pengelolaan data dalam jumlah besar, serta Apache Spark untuk *preprocessing* dan analisis sentimen. Dengan menerapkan pendekatan ini, penelitian diharapkan dapat memberikan wawasan yang lebih mendalam mengenai persepsi pengguna terhadap Layanan Gojek serta memberikan rekomendasi yang relevan bagi penyedia layanan transportasi *online*.

## 2. Metode Penelitian

Penelitian ini dimulai dengan tahap pengumpulan data ulasan mengenai transportasi online, kemudian dilanjutkan dengan *preprocessing* untuk membersihkan dan menyiapkan data. Setelah itu data diberi label sentimen sesuai kategori sebelum dilakukan pemodelan sentimen menggunakan algoritma *Logistic Regression*. Tahapan terakhir adalah evaluasi dan visualisasi hasil untuk menilai kinerja model yang diterapkan. Alur penelitian secara lengkap dapat dilihat pada Gambar 2.




Gambar 2. Alur Penelitian

### 2.1. Pengumpulan Data

Data ulasan pengguna Gojek diperoleh dari dua sumber utama, yaitu *dataset* GojekAppReview yang tersedia pada repositori Kaggle dan data yang diambil langsung dari Google Play Store menggunakan teknik web *scraping*. *Dataset* Kaggle berisi 225.043 ulasan yang dikumpulkan dalam periode November 2021 hingga Februari 2024, sedangkan data hasil *scraping* dari Google Play Store mencakup 1.655.069 ulasan yang dikumpulkan dari Mei 2019 hingga November 2024.

Teknik *scraping* dilakukan di Google Colab menggunakan metode *HDM Vision Google Play Scraper*, dengan bantuan *library Google Play Scraper*. Parameter yang diterapkan dalam proses pengambilan data mencakup aplikasi Gojek dengan ID 'com.gojek.app', ulasan berbahasa Indonesia (lang='id'), serta ulasan dari pengguna di Indonesia (country='id'). Data yang dikumpulkan difokuskan pada ulasan yang dianggap paling relevan (sort=Sort.MOST_RELEVANT) tanpa adanya pembatasan skor ulasan tertentu (filter_score_with=None). Periode pengambilan data mengikuti ketersediaan ulasan terbaru di Play Store, sehingga dataset yang diperoleh mencerminkan pengalaman pengguna terkini terhadap layanan Gojek.

### 2.2. *Preprocessing* Data

Dataset yang telah dikumpulkan diunggah ke dalam *Hadoop Distributed File System* (HDFS) untuk pengelolaan data secara terdistribusi. Sistem Hadoop digunakan untuk memastikan efisiensi penyimpanan dan akses terhadap dataset berukuran besar. Untuk proses *preprocessing* data dilakukan menggunakan Apache Spark. Tahapan *preprocessing* data meliputi:

DOI: 10.54259/satesi.v5i1.4051

a. Pembersihan data (*cleaning data*), yaitu menghapus karakter khusus seperti tanda baca, angka yang tidak relevan, URL, *mention* (@username), dan *hashtag* (#hashtag). Misalnya, sebuah ulasan seperti "@gojek Kenapa saldo saya tiba-tiba hilang?!!! #gojek" akan dibersihkan menjadi "Kenapa saldo saya tiba tiba hilang" dengan menggunakan *library* re dan PySpark.
b. Normalisasi, yaitu mengubah kata-kata yang tidak baku menjadi bentuk yang lebih standar, seperti mengubah "gk" menjadi "tidak". Contohnya, kalimat "gk ngerti nih driver kok lama" setelah normalisasi akan menjadi "tidak mengerti ini driver kok lama". Proses ini dilakukan menggunakan *library* NLTK.
c. Tokenisasi, yaitu proses memisahkan teks ulasan menjadi kata-kata tunggal sebagai unit individu agar dapat dianalisis lebih baik. Sebagai contoh, teks "Driver datangnya lama sekali" setelah melalui proses tokenisasi akan diubah menjadi ["Driver", "datang", "lama", "sekali"] menggunakan NLTK.word_tokenize.
d. Eliminasi *stopwords (stopword removal)*, yaitu proses menghapus kata-kata umum yang tidak memiliki pengaruh signifikan dalam analisis sentimen, seperti "dan", "atau", dan "yang". Penghapusan *stopwords* dilakukan dengan *library* NLTK.corpus.stopwords, yang akan menghilangkan kata-kata tidak penting. Sebagai contoh, teks "saya suka aplikasi ini karena mudah" akan berubah menjadi ["suka", "aplikasi", "mudah"] setelah *stopwords* dihapus.
e. Penanganan data hilang, yaitu menghapus ulasan kosong atau tidak relevan untuk menjaga kualitas data[17]. Proses ini dilakukan menggunakan *library* PySpark dengan fungsi dropna() digunakan untuk menghapus baris yang memiliki nilai kosong, dan drop_duplicates() digunakan untuk menghapus data duplikat guna menjaga kualitas dataset.

Penggunaan Spark memastikan proses ini berjalan secara paralel, sehingga efisien untuk dataset berukuran besar.

### 2.3. Pelabelan Data

Setelah *preprocessing* selesai, dataset diberi label sentimen untuk mengklasifikasi ulasan ke menjadi 3 kategori yaitu positif, negatif dan netral. Pelabelan dilakukan dengan pendekatan semi-otomatis, pelabelan ini diberikan berdasarkan rating pengguna, di mana ulasan dengan rating rendah dianggap negatif, rating sedang dianggap netral, dan rating tinggi dianggap positif. Untuk memastikan konsistensi antara rating dan isi teks ulasan, dilakukan validasi label dengan meninjau ulang beberapa data secara manual. Ulasan yang tidak memiliki rating atau memiliki isi yang tidak konsisten dengan rating, pelabelan dilakukan secara manual untuk memastikan akurasi. Dataset berlabel ini kemudian digunakan dalam proses pelatihan model.

Table 1. Kriteria pelabelan

| *Score* | Label |
|---|---|
| 1-2 | Negatif |
| 3 | Netral |
| 4-5 | Positif |

Sebagai contoh, ulasan "Aplikasi sering error, saldo Gopay hilang!" dengan *score* 1 diberi label "Negatif", sedangkan "Lumayan, tapi kadang lambat" dengan *score* 3 diberi label "Netral", dan "Gojek sangat membantu, selalu cepat dan tepat!" dengan *score* 5 diberi label "Positif".

### 2.4. Pemodelan Sentimen

Setelah tahap *preprocessing* dan pelabelan, data digunakan untuk membangun model analisis sentimen. Sebelum melakukan analisis sentimen, data teks diubah ke dalam bentuk numerik serta menggunakan TF-IDF untuk mengidentifikasi kata-kata yang relevan untuk analisis sentimen.

Untuk analisis sentimen, penelitian ini menggunakan algoritma *Logistic Regression*, yang dipilih karena kesederhanaannya, efektivitasnya dalam menangani data teks, serta kemampuannya memberikan probabilitas *output* yang terkalibrasi dengan baik. *Logistic Regression* juga lebih efisien dibandingkan *Support Vector Machine* (SVM) untuk *dataset* berukuran besar dan memiliki interpretasi yang lebih sederhana dibandingkan metode kompleks seperti *Random Forest.*

Pada tahap pelatihan, *dataset* dibagi menjadi data pelatihan (80%) dan data pengujian (20%) menggunakan fungsi randomSplit() dengan *seed* = 0 untuk menjamin hasil yang dapat direproduksi. Validasi model dilakukan dengan *Cross Validation* menggunakan CrossValidator, yang bertujuan untuk mencari kombinasi parameter terbaik selama proses pelatihan. Teknik validasi ini membagi data pelatihan menjadi beberapa *subset (folds)* dan menguji model pada setiap *subset* secara bergantian untuk memastikan hasil yang lebih stabil dan menghindari *overfitting*.

### 2.5. Evaluasi dan Visualisasi

Evaluasi model dilakukan dengan menggunakan *Logistic Regression* yang disertai dengan *hyperparameter tuning*. *Hyperparameter tuning* dilakukan dengan menggunakan CrossValidator dan parameter grid yang terdiri dari dua *hyperparameter* utama. Yang pertama adalah regParam (parameter regulasi) dengan nilai [0.01, 0.1, 1.0], yang berfungsi mengontrol kompleksitas model dan mencegah terjadinya *overfitting*. Nilai kecil (0.01) memungkinkan model menangkap pola kompleks, nilai menengah (0.1) memberikan keseimbanan antara fleksibilitas dan regularisasi, serta nilai besar (1.0) efektif untuk *dataset* dengan *noise* tinggi. Parameter kedua adalah elasticNetParam (parameter campuran ElasticNet) dengan nilai [0.0, 0.5, 1.0], yang menentukan kombinasi antara regularisasi L1 (Lasso) dan L2 (Ridge). Di mana nilai 0.0 (L2)

DOI: 10.54259/satesi.v5i1.4051

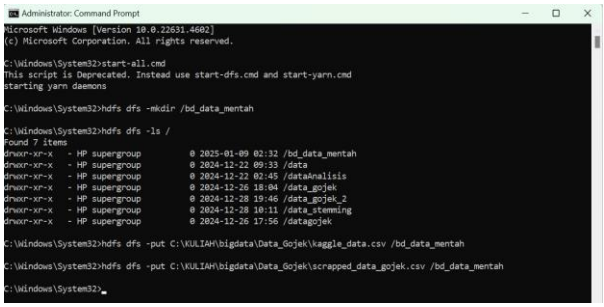
mempertahankan semua fitur dan mengurangi *overfitting* pada fitur yang berkorelasi, nilai 0.5 memberikan keseimbangan antara seleksi fitur dan stabilitas model, serta 1.0 (L1) menghilangkan fitur yang tidak relevan agar model lebih sederhana dan interpretative. Rentang ini dipilih untuk memastikan model mendapatkan kombinasi terbaik antara akurasi dan generalisasi.

Visualisasi dengan menggunakan *Word Cloud* dilakukan untuk menggambarkan kata-kata yang paling sering muncul dalam ulasan pengguna. *Word Cloud* dibuat untuk setiap kategori sentimen, yaitu positif, negatif, dan netral.

## 3. Hasil dan Pembahasan

### 3.1. Menyiapkan Data

Data ulasan aplikasi Gojek yang telah diperoleh dari Kaggle dan hasil *scraping* di Play Store disiapkan dengan menggunakan bantuan Hadoop dan PySpark, serta kode editor yang digunakan adalah Jupyter Notebook. Langkah pertama dalam menyiapkan data adalah dengan mengunggah data ke Hadoop.




Gambar 3. Proses Pengunggahan data Ulasan ke Hadoop

Setelah data berhasil diunggah ke Hadoop, data Kaggle dan data hasil *scraping* dari Play Store disatukan dengan menggunakan pyspark yang dijalankan di Jupyter Notebook. Sebelum menyatukan data, atribut yang tidak dibutuhkan seperti *username* dihilangkan, dan menyiasakan atribut *content*, *score*, at, dan appVersion pada data Kaggle. Sementara itu, untuk data hasil *scraping* dari Play Store atribut yang dihapus adalah *reviewId*, *username*, *userImage*, *thumbsUpCount*, *reviewCreatedVersion*, *replyContent*, *repliedAt* dan menyisakan atribut *content*, *score*, at, dan appVersion.

| content | score | at | appVersion |
|---|---|---|---|
| Maaf gw beri bint... | 2 | 2022-11-10 09:57:23 | 4.9.3 |
| Kok gada mode pil... | 1 | 2024-01-29 11:18:38 | 4.81.2 |
| Pedaeh urung di u... | 4 | 2024-01-24 02:20:56 | 4.81.2 |
| ok sip | 3 | 2024-01-22 11:05:23 | 4.81.2 |
| Gopaylater tiba2 ... | 1 | 2024-02-01 18:09:40 | 4.81.2 |

Gambar 3. Pemilihan atribut penting pada data Sebelum tahap *Preprocessing*

Selanjutnya, untuk masing-masing data disesuaikan atribut dan tipe datanya agar dapat disatukan. Setelah itu, barulah kedua data tersebut dapat disatukan.

| content | score | at | appVersion |
|---|---|---|---|
| Maaf gw beri bint... | 2 | 2022-11-10 09:57:23 | 4.9.3 |
| Kok gada mode pil... | 1 | 2024-01-29 11:18:38 | 4.81.2 |
| Pedaeh urung di u... | 4 | 2024-01-24 02:20:56 | 4.81.2 |
| ok sip | 3 | 2024-01-22 11:05:23 | 4.81.2 |
| Gopaylater tiba2 ... | 1 | 2024-02-01 18:09:40 | 4.81.2 |
| Selalu bisa di an... | 5 | 2024-01-21 23:35:50 | 4.81.1 |
| Aplikasi yang san... | 5 | 2024-01-21 17:19:01 | 4.81.1 |
| Semakin kesini se... | 5 | 2024-01-19 11:58:00 | 4.81.1 |
| Ok | 5 | 2024-01-15 09:14:18 | 4.81.1 |
| Terimakasih sanga... | 4 | 2024-01-15 00:09:35 | 4.81.1 |
| sejauh ini masih ... | 4 | 2024-01-17 09:06:34 | 4.81.1 |
| Keren banget cihuy | 5 | 2024-01-17 08:25:44 | 4.81.1 |
| Makin KAMUKO | 5 | 2024-01-01 04:21:00 | 4.80.4 |
| Bunga pinjam ting... | 1 | 2024-01-04 10:29:22 | 4.80.4 |
| Murah pool | 5 | 2024-01-04 01:35:12 | 4.80.4 |
| Ok | 5 | 2024-01-04 05:32:23 | 4.80.4 |
| Kenapa akun saya ... | 5 | 2024-01-05 04:59:51 | 4.80.4 |
| terlalu mahalgak ... | 1 | 2024-01-05 10:53:02 | 4.80.4 |
| pelayanan bagus w... | 4 | 2024-01-02 17:13:29 | 4.80.4 |
| Drivernya lama | 2 | 2023-12-26 15:10:48 | 4.80.4 |

Gambar 4. Penggabungan dataset dari Kaggle dan Play Store

Jumlah data menjadi 1.880.112 ulasan setelah proses penggabungan. Data tersebut kemudian dikirim kembali ke Hadoop untuk proses lebih lanjut.

### 3.2. *Preprocessing* Data

Tahap ini dimulai dengan menghapus data yang kosong maupun duplikat di dalam data. Selanjutnya, data dibersihkan dari elemen-elemen yang tidak relevan. Misalnya dengan menghapus URL, *mention*, dan *hashtag*. Selain itu, karakter non-alfanumerik yang tidak diperlukan, seperti simbol atau tanda baca tertentu juga dihilangkan. Spasi berlebihan yang mungkin terdapat pada data, termasuk di awal dan di akhir teks juga dihapus agar format data menjadi lebih rapi.

| content | score | at | appVersion |
|---|---|---|---|
| A | 5 | 2020-03-21 09:35:40 | 3.46.1 |
| bagussssssssss | 5 | 2020-11-21 17:35:18 | 4.8.1 |
| b aja seh | 4 | 2021-06-08 05:16:09 | 4.20.1 |
| Karya anak bangsa Mantaaaap | 5 | 2019-08-31 23:11:53 | 3.27.0 |
| lagend di atas kasur itu c | 5 | 2019-07-18 00:18:46 | 3.32.1 |

Gambar 5. Setelah Penghapusan Karakter Non-Alfanumerik, URL, Mention, dan Hashtag

Setelah dilakukan pembersihan, semua huruf kapital pada kolom konten diubah menjadi huruf kecil. Transformasi ini dilakukan untuk menghindari perbedaan yang tidak relevan., misalnya antara “Baru” dan “baru,” yang sebenarnya memiliki arti yang sama tetapi dikenali berbeda oleh sistem.

Gambar 6. Mengubah Teks Ulasan menjadi Huruf Kecil

Langkah selanjutnya adalah normalisasi data. Pada proses ini dilakukan koreksi kesalahan ejaan atau *typo* yang sering terjadi dalam teks ulasan. Misalnya, kata-kata yang salah penulisan seperti “alesan” akan dikoreksi menjadi “alasan” agar teks lebih seragam.

DOI: 10.54259/satesi.v5i1.4051

Setelah dinormalisasi, teks dipecah menjadi unit-unit kecil berupa kata dengan proses tokenisasi. Tokenisasi memungkinkan algoritma menganalisis setiap kata dalam teks secara individual, seperti memisahkan "lambat sekali sinyalnya" menjadi "[lambat, sekali, sinyalnya]."

| content | tokenisasi |
|---|---|
| lambat sekali sinyalnya | [lambat, sekali, sinyalnya] |
| mengagumkan | [mengagumkan] |
| keren | [keren] |
| mantap | [mantap] |
| update terus jadi malas | [update, terus, jadi, malas] |

only showing top 5 rows

Gambar 7. Tokenisasi Teks Ulasan

Pembersihan lebih lanjut dilakukan dengan menghilangkan kata-kata yang tidak penting, seperti *stopwords*. Misalnya, kata "dan," "tidak," "di," atau "ini," dianggap tidak memberikan nilai tambah pada analisis dan akan dihapus untuk mengurangi *noise* dalam data.

| tokenisasi | stopwords |
|---|---|
| [lambat, sekali, sinyalnya] | [lambat, sekali, sinyalnya] |
| [mengagumkan] | [mengagumkan] |
| [keren] | [keren] |
| [mantap] | [mantap] |
| [update, terus, jadi, malas] | [update, malas] |

only showing top 5 rows

Gambar 8. Menghapus *Stopwords* dari Teks Ulasan

Hasil pembersihan pada kolom *stopwords* yang tadinya masih dalam bentuk *array* diubah menjadi *string*. Hal ini dilakukan agar data dapat disimpan dalam bentuk CSV dan lebih mudah digunakan untuk analisis selanjutnya. Setelah itu, dilakukan pemilihan kolom yang relevan untuk analisis sentimen. Kolom yang dipilih meliputi *score*, *at* yang menunjukkan tanggal ulasan dibuat, *appVersion* untuk mengetahui versi aplikasi, dan *stopword_string* sebagai hasil akhir dari *preprocessing* data.

| score | at | appVersion | stopwords_string |
|---|---|---|---|
| 5 | 2020-03-21 09:35:40 | 3.46.1 | a |
| 5 | 2020-11-21 17:35:18 | 4.8.1 | bagus |
| 4 | 2021-06-08 05:16:09 | 4.20.1 | biasa |
| 5 | 2019-08-31 23:11:53 | 3.27.0 | karya anak bangsa... |
| 5 | 2019-07-18 00:18:46 | 3.32.1 | legenda atas kasur c |
| 2 | 2023-08-05 13:27:32 | 4.69.2 | mxk xak |
| 5 | 2020-02-25 06:59:56 | 3.39.2 | adil harus samaun... |
| 4 | 2020-08-13 02:20:23 | 3.40.2 | biasa |
| 1 | 2020-12-28 03:37:27 | 4.10.3 | beri ranting kare... |
| 5 | 2018-09-26 05:59:11 | 3.13.2 | sangat mndukung k... |
| 5 | 2021-03-06 13:09:28 | 4.8.1 | something went wr... |

Gambar 9. Pemilihan kolom Setelah Proses Preprocessing data untuk Analisis

Jumlah data awal sebelum *preprocessing* adalah sebanyak 1.880.112 ulasan. Setelah melalui proses *preprocessing* jumlah data akhir yang digunakan dalam analisis adalah 731.946

### 3.3. Pelabelan Data

Data diberi label untuk membantu klasifikasi data ke dalam kategori tertentu berdasarkan atribut. Pelabelan bertujuan untuk mengidentifikasi apakah suatu data termasuk ke dalam kategori positif, negatif, atau netral. Pada penelitian ini, data diberi label berdasarkan atribut *score*. Jika *score* lebih besar sama dengan satu atau *score* yang kurang dari atau sama dengan dua maka akan diberi label negatif, jika *score* sama dengan tiga akan diberi netral, dan apabila *score* lebih besar atau sama dengan empat dan *score* kurang dari sama dengan lima maka akan diberi label positif.

| score | at | appVersion | stopwords_string | label |
|---|---|---|---|---|
| 5 | 2020-03-21 09:35:40 | 3.46.1 | a | positif |
| 5 | 2020-11-21 17:35:18 | 4.8.1 | bagus | positif |
| 4 | 2021-06-08 05:16:09 | 4.20.1 | biasa | positif |
| 5 | 2019-08-31 23:11:53 | 3.27.0 | karya anak bangsa... | positif |
| 5 | 2019-07-18 00:18:46 | 3.32.1 | legenda atas kasur c | positif |
| 2 | 2023-08-05 13:27:32 | 4.69.2 | mxk xak | negatif |
| 5 | 2020-02-25 06:59:56 | 3.39.2 | adil harus samaun... | positif |
| 4 | 2020-08-13 02:20:23 | 3.40.2 | biasa | positif |
| 1 | 2020-12-28 03:37:27 | 4.10.3 | beri ranting kare... | negatif |
| 5 | 2018-09-26 05:59:11 | 3.13.2 | sangat mndukung k... | positif |
| 5 | 2021-03-06 13:09:28 | 4.8.1 | something went wr... | positif |
| 4 | 2018-02-28 05:31:52 | 3.0.1 | peta perjalanan a... | positif |
| 1 | 2022-01-24 09:53:39 | 4.35.1 | knpa gopay masih ... | negatif |

Gambar 10. Pelabelan Data Berdasarkan Score Ulasan

### 3.4. Pemodelan Sentimen

Sebelum melakukan analisis sentimen, data diubah ke dalam bentuk numerik. Hal ini penting untuk dilakukan karena algoritma *machine learning* dirancang untuk bekerja dengan angka, bukan dengan teks mentah[18]. Proses ini dilakukan dengan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Kemudian, model *Logistic Regression* diimplementasikan untuk klasifikasi sentimen. Dataset dibagi menjadi data pelatihan (80%) dan data pengujian (20%) menggunakan fungsi randomSplit() dengan seed = 0 untuk menjamin hasil yang dapat direproduksi.

Kemudian, model *Logistic Regression* dikonfigurasi dengan parameter dasar maxIter = 10, yang membatasi jumlah iterasi maksimum dalam proses optimasi. Untuk menemukan parameter yang paling optimal, dilakukan penyesuaian *hyperparameter* melalui teknik *Cross Validation* dengan dua parameter utama: regParam (parameter regularisasi) yang diuji dengan nilai [0.01, 0.1, 1.0] untuk mengontrol kompleksitas model dan mencegah terjadinya overfitting, serta elasticNetParam (parameter campuran ElasticNet) yang diuji dengan nilai [0.0, 0.5, 1.0] untuk mencapai keseimbangan antara regularisasi L1 dan L2.


Gambar 11. Proses Konfigurasi dan Tuning Model *Logistic Regression* pada PySpark

DOI: 10.54259/satesi.v5i1.4051

*Cross Validation* dilakukan dengan membagi data training menjadi k-fold dan menguji kombinasi parameter pada setiap *fold* untuk menemukan parameter terbaik[19].

Hasil evaluasi menunjukkan bahwa model *Logistic Regression* setelah *hyperparameter tuning* mendapatkan akurasi sebesar 0,80 atau 80%. Ini menunjukkan bahwa model memiliki kinerja yang cukup baik, dalam mengklasifikasi sentimen ulasan pengguna aplikasi Gojek ke dalam kategori negatif, netral, dan positif. Penggunaan PySpark memungkinkan proses *training* dan evaluasi dapat berjalan dengan efisien meskipun dengan dataset berukuran besar.

3.5. Visualisasi Data

Visualisasi data dibuat berdasarkan hasil TF-IDF, dimana kata-kata yang memiliki bobot lebih tinggi dalam perhitungan TF-IDF sering kali mencerminkan opini atau sentimen pengguna yang lebih signifikan, sehingga memberikan informasi tambahan yang berguna pada proses analisis[20].

Setelah data dikonversi ke dalam bentuk numerik, Langkah selanjutnya adalah membuat visualisasi berupa *wordcloud*. Visualisasi ini bertujuan untuk memberikan gambaran tentang kata-kata yang paling sering muncul dalam data ulasan[21].

Pada *wordcloud* sentimen negatif, keluhan pengguna terhadap layanan Gojek terutama berpusat pada performa aplikasi dan fitur dompet digital, Gopay. Kata-kata yang sering muncul seperti "aplikasi", "gojek", "saldo", dan "gopay" menunjukkan adanya masalah signifikan terkait fungsi aplikasi dan sistem pembayaran digital. Penggunaan kata "tolong" yang sering muncul mencerminkan tingginya permintaan bantuan dari pengguna, sementara kata "karena" menunjukkan bahwa pengguna memberikan alasan khusus atas ketidakpuasan mereka. Kata "masih" yang muncul dalam konteks negatif dapat mengindikasikan adanya masalah yang belum terselesaikan atau terus berlanjut. Perhatikan gambar 12.


Gambar 12. *Wordcloud* Ulasan Negatif

Secara keseluruhan, ulasan negatif menandakan perlunya perbaikan pada kinerja aplikasi, kejelasan informasi, serta kecepatan respons saat menghadapi masalah pada fitur keuangan. Dalam upaya mengatasi isu tersebut, Gojek dapat memperbaiki infrastruktur *backend* untuk meningkatkan stabilitas sistem transaksi Gopay dan mengurangi kemungkinan transaksi gagal. Selain itu, transparansi saldo dapat ditingkatkan dengan penerapan pencatatan transaksi secara *real-time* serta penyediaan fitur pelacakan riwayat saldo yang lebih detail. Selain itu, implementasi notifikasi proaktif mengenai gangguan atau keterlambatan transaksi akan membantu mengurangi kebingungan pengguna sekaligus meningkatkan kepercayaan mereka terhadap layanan.

Pada word cloud dengan sentimen netral, pengguna memberikan ulasan yang tidak sepenuhnya positif maupun negatif. Kata "*driver*" menjadi salah satu yang paling sering muncul, menunjukkan bahwa pengalaman dengan pengemudi merupakan aspek utama yang diperhatikan pelanggan. Selain itu, kata "aplikasi" juga banyak muncul menunjukkan perhatian pengguna terhadap fitur dan performa aplikasi Gojek yang masih memerlukan peningkatan. Kata "Gojek" sering disampaikan dalam konteks yang netral, lebih sebagai sebutan merek tanpa emosi yang kuat. Kemunculan kata "masih" dan "lama" menunjukkan bahwa pelanggan cenderung membandingkan layanan saat ini dengan pengalaman sebelumnya, yang mencerminkan adanya harapan terhadap peningkatan kualitas. Selain itu, penggunaan kata "tolong" yang cukup sering menandakan bahwa meskipun *feedback* tersebut tidak selalu berbentuk keluhan langsung, pengguna tetap berharap adanya peningkatan atau perbaikan dalam layanan. Perhatikan gambar 13.


Gambar 13. *Wordcloud* Ulasan Netral

Ulasan netral ini dapat menjadi masukan berharga bagi Gojek dalam mengevaluasi aspek layanan yang masih perlu ditingkatkan. Mengingat aspek pengemudi menjadi perhatian utama, meningkatkan pelayanan dari pengemudi sangat berpengaruh terhadap kepuasan pelanggan. Oleh karena itu, Gojek dapat memperkuat program pelatihan dan evaluasi berkala bagi mitra

DOI: 10.54259/satesi.v5i1.4051

driver untuk meningkatkan standar layanan. Di sisi lain, perhatian terhadap fitur dan stabilitas serta fungsionalitas sistem menunjukkan perlunya optimalisasi performa aplikasi serta peningkatan pengalaman pengguna (UX) dengan antarmuka yang lebih intuitif dan responsif. Selain itu, pelanggan menginginkan perbaikan layanan yang lebih cepat. Oleh karena itu, Gojek perlu lebih proaktif dalam mengomunikasikan pembaruan layanan serta memberikan transparansi terkait proses perbaikan yang sedang dilakukan.

Pada *wordcloud* yang menggambarkan sentimen positif, istilah seperti “sangat membantu” dan “membantu sekali” muncul dengan ukuran yang paling besar, mencerminkan rasa syukur dan kepuasan pengguna terhadap layanan Gojek. Ungkapan "terima kasih" yang sering muncul mencerminkan tingkat kepuasan dan apresiasi yang tinggi dari pengguna. Frasa "aplikasi Gojek" yang muncul dalam konteks positif menandakan bahwa banyak pengguna memiliki pengalaman yang memuaskan dengan platform. Perhatikan gambar 14.


Gambar 14. *Wordcloud* Ulasan Positif

*Word cloud* ini menunjukkan bahwa meskipun terdapat beberapa keluhan, banyak pengguna merasa terbantu oleh layanan Gojek. Untuk mempertahankan kepuasan pelanggan, Gojek dapat terus meningkatkan fitur-fiturnya dengan memperhatikan masukan negatif.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negatif | 0.83 | 0.94 | 0.88 | 83091.0 |
| Positif | 0.75 | 0.70 | 0.72 | 38585.0 |
| Netral | 0.23 | 0.03 | 0.06 | 9367.0 |
| accuracy | 0.80 | 0.80 | 0.80 | 0.8 |
| macro avg | 0.60 | 0.56 | 0.55 | 131043.0 |
| weighted avg | 0.76 | 0.80 | 0.77 | 131043.0 |

Gambar 15. Hasil Evaluasi Model

Hasil evaluasi model ditampilkan pada gambar 16, yang menunjukkan evaluasi model berupa *precision*, *recall*, dan *f1-score* untuk masing-masing kelas: Negatif, Positif, dan Netral. Model mencapai akurasi keseluruhan sebesar 80%, sebuah hasil yang signifikan mengingat ketidakseimbangan distribusi data antar kelas. Model menunjukkan performa yang sangat baik dalam mengklasifikasikan kelas Negatif, dengan *precision* sebesar 0.83 dan *recall* mencapai 0.94. Hal ini menunjukkan bahwa model mampu mengidentifikasi sebagian besar data negatif dengan tingkat kesalahan yang rendah. Sementara itu, kelas Positif menunjukkan performa yang seimbang dengan *precision* 0.75 dan *recall* 0.70, menandakan bahwa model cukup baik dalam membedakan sentimen positif. Namun, performa kelas Netral tergolong rendah, dengan *recall* hanya sebesar 0.03. Hal ini dapat dikaitkan dengan distribusi data yang tidak seimbang, di mana jumlah data netral jauh lebih sedikit dibandingkan kelas lainnya. Meskipun demikian, model tetap mampu mencapai keseimbangan performa global yang cukup baik, sebagaimana ditunjukkan oleh nilai *weighted average f1-score* sebesar 0.77. Untuk mengatasi ketidakseimbangan data dan meningkatkan performa model, penelitian ini telah menerapkan teknik *hyperparameter tuning* menggunakan *CrossValidator* dan parameter *grid*. Dengan demikian, model tetap mampu memberikan hasil klasifikasi yang optimal dan menjaga keseimbangan performa antar kelas.

Hasil penelitian ini dibandingkan dengan studi sebelumnya yang menggunakan metode berbeda. Penelitian yang dilakukan oleh Saripah Aini Pohan mengenai analisis sentimen pada aplikasi Maxim menggunakan algoritma Random Forest memperoleh akurasi 64%(Aini Pohan & Hasyifah Sibarani,2024)[15]. Sementara itu, penelitian Melia yang menerapkan Random Forest untuk analisis sentimen pengguna Gojek dan Grab melalui media sosial Twitter mencapai akurasi 76,25%(Irawan & Nurdiawan, 2023)[16]. Dibandingkan dengan penelitian ini, menggunakan metode *Logistic Regression* yang dioptimalkan dengan Hadoop dan PySpark berhasil mencapai akurasi 80%, menunjukkan peningkatan performa dalam klasifikasi sentimen ulasan aplikasi transportasi online. Keunggulan utama penelitian ini adalah penggunaan big data *processing*, yang memungkinkan analisis data dalam jumlah besar lebih cepat dan efisien dibandingkan metode sebelumnya. Selain itu, penerapan *hyperparameter tuning* membantu meningkatkan keseimbangan performa model meskipun terdapat ketidakseimbangan data. Namun, penelitian ini masih memiliki beberapa keterbatasan. Salah satunya adalah rendahnya *recall* pada sentimen netral akibat jumlah data ulasan netral yang lebih sedikit. Selain itu, *Logistic Regression* masih memiliki keterbatasan dalam memahami konteks kalimat yang kompleks, seperti sarkasme atau opini yang ambigu. Oleh karena itu, untuk penelitian selanjutnya, disarankan menggunakan model berbasis *deep learning*, seperti BERT atau Transformer, yang lebih mampu menangkap makna kontekstual dalam ulasan pengguna.

Analisis sentimen yang dilakukan memiliki keterkaitan erat dengan berbagai tantangan yang dihadapi Gojek, khususnya dalam meningkatkan kualitas layanan dan pengalaman pelanggan. Banyaknya ulasan negatif yang terdeteksi dalam analisis ini mengindikasikan

DOI: 10.54259/satesi.v5i1.4051

adanya beberapa aspek yang perlu diperbaiki, seperti stabilitas aplikasi, kendala dalam sistem pembayaran digital, serta kualitas layanan dari mitra pengemudi. Oleh karena itu, hasil penelitian ini dapat menjadi acuan bagi Gojek dalam menyusun strategi peningkatan layanan serta dapat lebih responsif dalam menangani keluhan pelanggan, mempercepat proses perbaikan layanan, dan secara keseluruhan meningkatkan kepuasan pengguna.

## 4. Kesimpulan

Penerapan Hadoop untuk analisis sentimen berbasis big data pada ulasan aplikasi Gojek berhasil mengolah dataset yang awalnya berjumlah 1.880.112 ulasan. Setelah melalui proses preprocessing, jumlah ulasan yang digunakan dalam analisis akhir menjadi 731.946. Hasil evaluasi model *Logistic Regression* menunjukkan performa yang baik, dengan akurasi keseluruhan sebesar 80%. Model memiliki kemampuan tinggi dalam mengklasifikasikan ulasan negatif (precision 0.83, recall 0.94), performa seimbang pada ulasan positif (precision 0.75, recall 0.70), namun performa rendah pada ulasan netral karena distribusi data yang tidak seimbang. Masalah ini diatasi melalui *hyperparameter tuning* menggunakan *CrossValidator* dan *parameter grid*, sehingga model tetap mampu menjaga keseimbangan performa secara keseluruhan. Penggunaan PySpark memungkinkan efisiensi dalam pemrosesan data skala besar, mempercepat proses training dan evaluasi model.

Hasil penelitian dapat dijadikan dasar untuk meningkatkan kualitas layanan, terutama dalam mengatasi permasalahan teknis yang sering dikeluhkan pengguna seperti perbaikan stabilitas aplikasi dan sistem pembayaran digital, peningkatan kualitas layanan *driver* untuk menjaga standar pelayanan yang konsisten. Selain itu, integrasi analisis sentimen dalam sistem *monitoring real-time* dapat membantu mendeteksi dan merespons permasalahan secara cepat, yang pada akhirnya dapat meningkatkan kepuasan pelanggan dan mendorong pertumbuhan bisnis Gojek secara keseluruhan serta meningkatkan strategi pemasaran agar memperkuat loyalitas pelanggan dan meningkatkan daya saing di pasar.

Untuk penelitian mendatang, analisis dapat diperluas dengan menambahkan dimensi temporal guna mengamati tren perubahan sentimen dari waktu ke waktu serta mengintegrasikan analisis sentimen dengan metrik bisnis lainnya untuk memberikan wawasan yang lebih menyeluruh dalam pengembangan layanan transportasi *online*. Selain itu, disarankan untuk menggunakan model berbasis deep learning, seperti BERT atau Transformer, guna meningkatkan akurasi dan pemahaman terhadap konteks dalam ulasan pengguna. Model ini lebih efektif dalam menangkap pola bahasa yang lebih kompleks, termasuk sarkasme atau opini tersirat yang sulit diidentifikasi oleh model berbasis machine learning konvensional.

## SUMBER RUJUKAN
### Referensi

[1] G. A. Pratiwi, R. M. Almakhsum, R. D. Setiyawati, A. P. Farahdila, and A. Zaki, "Kontestasi Start-up Ojek Online di Indonesia: Strategi Promosi Digital Gojek, Grab, Indriver, dan Maxim," *OIKONOMIKA : Jurnal Kajian Ekonomi dan Keuangan Syariah*, vol. 5, no. 1, pp. 64–79, Aug. 2024, doi: 10.53491/oikonomika.v5i1.955.

[2] R. Wahyudi *et al.*, "Analisis Sentimen pada review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *JURNAL INFORMATIKA*, vol. 8, no. 2, pp. 200–207, Sep. 2021, doi: https://doi.org/10.31294/ji.v8i2.9681.

[3] M. Syaharani, "Moda Transportasi Online Menjadi Pilihan Masyarakat, Apa Alasannya?," GoodStats. [Online]. Available: https://goodstats.id/article/moda-transportasi-online-semakin-menjadi-favorit-masyarakat-apa-alasannya-BFsC9

[4] V. K. S. Que, A. Iriani, and H. D. Purnomo, "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* |, vol. 9, no. 2, pp. 162–170, May 2020, doi: https://doi.org/10.22146/jnteti.v9i2.102.

[5] N. Yurindera, "Dampak Kepercayaan Terhadap Loyalitas Pelanggan Melalui Kepuasan pada Layanan Transportasi Ojek Online," *Jurnal Esensi Infokom*, vol. 8, no. 2, pp. 41–47, Oct. 2024, doi: https://doi.org/10.55886/infokom.v8i2.931.

[6] R. A. Fauzi, I. Cholissodin, and B. Rahayudi, "Pemanfaatan Spark untuk Analisis Sentimen Mengenai Netralitas Berita dalam Membahas Pemilu Presiden 2019 Menggunakan Metode Naïve Bayes Classifier," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 3, pp. 1070–1077, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[7] M. N. Fadillah and D. Bernadisman, "Peranan Ojek Online dalam Meningkatkan Omzet UMKM dan Pertumbuhan Ekonomi Digital Indonesia," *JUARA: Jurnal Pengabdian Kepada Masyarakat*, vol. 1, no. 1, pp. 32–35, 2023, Accessed: Jan. 15, 2025. [Online]. Available: https://jurnas.saintekmu.ac.id/index.php/juara/article/view/57

[8] N. A. Amalia, I. T. Utami, and Y. Wilandari, "ANALISIS SENTIMEN KEBIJAKAN PENYELENGGARA SISTEM ELEKTRONIK LINGKUP PRIVAT MENGGUNAKAN PENALIZED LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE," *Jurnal Gaussian*, vol. 12, no. 4, pp. 560–569, Jul. 2024, doi: 10.14710/j.gauss.12.4.560-569.

[9] I. Rahmawati and T. R. Fitriani, "Analisis Sentimen Menggunakan Algoritma Logistic Regression Pada Penerbangan Lion Air berdasarkan Ulasan Pengguna Platform Online," *Jejaring Penelitian dan Pengabdian Masyarakat (JPPM)*, vol. 1, no. 1, pp. 11–16, Aug. 2023, doi: https://doi.org/10.58776/jriti.v1i1.60.

[10] K. D. Indarwati and H. Februariyanti, "ANALISIS SENTIMEN TERHADAP KUALITAS PELAYANAN APLIKASI GO-JEK MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER," *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, vol. 10, no. 10, Mar. 2023, doi: https://doi.org/10.35957/jatisi.v10i1.2643.

[11] Junaedi, A. Hendra Gunawan, V. Kuswanto, and Jonathan, "Eksplorasi Algoritma Support Vector Machine untuk Analisis Sentimen Destinasi Wisata di Indonesia," *Bit-Tech (Binary Digital - Technology)*, vol. 7, no. 2, pp. 323–330, Dec. 2024, doi: 10.32877/bt.v7i2.1810.

[12] A. Gugun, B. R. Sanjaya, F. Rahmadani, and J. C. Key, "LITERATURE REVIEW: Penggunaan Support Vector Machine (SVM) untuk Klasifikasi Penyakit Lambung," *Buletin Ilmiah Ilmu Komputer Dan Multimedia (BIIKMA)*, vol. 2, no. 3, pp. 546–549, Oct. 2024, doi: 10.32493/jtsi.v7i3.42254.

[13] I. Permatahati, M. N. Perdana, N. Apriadi, T. P. Amanda, and Z. Maharani, "Analisis Dataset TOP 1000 IMDb Movies Menggunakan Hadoop," *Journal of Network and Computer*, vol. 2, no. 2, pp. 23–36, 2023, [Online]. Available: https://jurnal.netplg.com/

DOI: 10.54259/satesi.v5i1.4051

[14] L. Arrahmando Romadhona, R. Febrianti, A. Winata, C. Putri Amanda, and R. Julia Erizka, "Hadoop-MapReduce Pada YARN Framework," *Journal of Network and Computer*, vol. 1, no. 2, pp. 91–101, 2022, [Online]. Available: https://jurnal.netplg.com/jnca

[15] S. Aini Pohan and F. Hasyifah Sibarani, "ANALISIS SENTIMEN TERHADAP APLIKASI MAXIM MENGGUNAKAN ALGORITMA RANDOM FOREST," *Journal of Science and Social Research*, no. 3, pp. 1201–1208, 2024, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[16] B. Irawan and O. Nurdiawan, "ANALISIS SENTIMEN TERHADAP PENGGUNA GOJEK DAN GRAB PADA MEDIA SOSIAL TWITTER MENGGUNAKAN RANDOM FOREST," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 5, pp. 3614–3618, 2023.

[17] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, pp. 406–414, Apr. 2021, doi: 10.30865/mib.v5i2.2835.

[18] D. A. Hamidah, R. Salkiawati, and R. Sari, "Analisis Sentimen Ulasan Customer Kopi TMLST Menggunakan Algoritma Naïve Bayes," *Journal of Students' Research in Computer Science*, vol. 5, no. 1, pp. 27–40, May 2024, doi: 10.31599/mrm89y71.

[19] W. Wijiyanto, A. I. Pradana, S. Sopingi, and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," *Jurnal Algoritma*, vol. 21, no. 1, pp. 239–248, May 2024, doi: 10.33364/algoritma/v.21-1.1618.

[20] A. N. Hasanah and B. N. Sari, "ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI JASA OJEK ONLINE MAXIM PADA GOOGLE PLAY DENGAN METODE NAÏVE BAYES CLASSIFIER," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, pp. 90–96, Jan. 2024, doi: 10.23960/jitet.v12i1.3628.

[21] J. A. Wibowo, V. C. Mawardi, and T. Sutrisno, "VISUALISASI WORD CLOUD HASIL ANALISIS SENTIMEN BERBASIS FITUR LAYANAN APLIKASI GOJEK DENGAN SUPPORT VECTOR MACHINE," *Jurnal Serina Sains, Teknik dan Kedokteran*, vol. 2, no. 1, pp. 61–70, Mar. 2024, doi: 10.24912/jsstk.v2i1.32058.

DOI: 10.54259/satesi.v5i1.4051